WSU-SmartEnv-NG Metrics[1]

The following metrics are used to evaluate a novelty-aware agent. The names correspond to the names used in the CSV spreadsheet outputs of the analysis script.

Name	Measure	Description
M1	$\overline{\text{FN}}_{\text{CDT}}$	Mean number of false negatives among CDTs. Lower is better.
M2	CDT%	% of CDTs among all trials
M2.1	FP%	% of trials with at least one false positive. Lower is better.
M3 (ONRP) unknown	$\sum P_{Post,\alpha}/\sum P_{Pre,\beta}$	Overall Novelty Reaction Performance (ONRP): Agent post-novelty vs. Baseline pre-novelty (novel episodes unknown)
M3.1 (INRP) unknown	$\sum_{i=1}^{m} P_{Post,\alpha}\Big/\sum P_{Pre,\beta}$	Initial Novelty Reaction Performance (INRP): Agent initial post-novelty vs. Baseline pre-novelty (novel episodes unknown)
M4 (ONRP) known	$\sum P_{Post,\alpha}/\sum P_{Pre,\beta}$	Overall Novelty Reaction Performance (ONRP): Agent post-novelty vs. Baseline pre-novelty (novel episodes known)
M4.1 (INRP) known	$\sum_{i=1}^{m} P_{Post,\alpha}\Big/\sum P_{Pre,\beta}$	Initial Novelty Reaction Performance (INRP): Agent initial post-novelty vs. Baseline pre-novelty (novel episodes known)
OPTI	$\frac{\sum P_{Post,\alpha}}{\sum P_{Post,\alpha} + \sum P_{Post,\beta}}$	Overall Performance Task Improvement (OPTI): Agent vs. Baseline post-novelty
IPTI	$\frac{\sum_{i=1}^{m} P_{Post,\alpha}}{\sum_{i=1}^{m} P_{Post,\alpha} + \sum_{i=1}^{m} P_{Post,\beta}}$	Initial Performance Task Improvement (IPTI): Agent vs. Baseline initial post-novelty
APTI	$\frac{\sum_{i=N_{post}-m}^{N_{post}} P_{Post,\alpha}}{\sum_{i=N_{post}-m}^{N_{post}} P_{Post,\alpha} + \sum_{i=N_{post}-m}^{N_{post}} P_{Post,\beta}}$	Asymptotic Performance Task Improvement (APTI): Agent vs. Baseline final post-novelty
AMOC	$\sum FN(FP\ rate)$	Area under the Activity Monitoring Operating Characteristic (AMOC) curve[2]. FN as a function of FP rate.

[1] These metrics were developed as part of the DARPA SAIL-ON program.

[2] https://pmc.ncbi.nlm.nih.gov/articles/PMC2815453/

NRM	$NRM = \frac{1}{N_{Trials}} \sum_{i=1}^{N_{Trials}} NRM(T_i)$ $NRM(T) = \frac{\left\|\mu\left(P_{Post,\alpha}\right) - \mu\left(P_{pre,\alpha}\right)\right\|}{\sigma\left(P_{pre,\alpha}\right)} < 2$	Novelty Robustness Measure (NRM): Percentage of trials for which the difference between the mean post-novelty and pre-novelty performance is less than 2 standard deviations of pre-novelty performance. Target agent α. Lower is better.
NRM_beta	Same as above, but replace α with β.	NRM for Baseline agent (β). Lower is better.
M2.2	TN/N_{Pre}	True negatives among pre-novelty episodes.
PRE_SOTA	$P_{Pre,\beta}$	Performance of Baseline agent pre-novelty.
PRE_TA2	$P_{Pre,\alpha}$	Performance of Target agent pre-novelty
POST_SOTA	$P_{Post,\beta}$	Performance of Baseline agent post-novelty.
POST_TA2	$P_{Post,\alpha}$	Performance of Target agent post-novelty.

Terminology:

- SOTA represents the Baseline Agent, which is assumed to be a state-of-the-art non-novelty-aware AI agent. The beta β symbol is also used to refer to this agent.
- TA2 represents the novelty-aware Target Agent. The alpha α symbol is also used to refer to this agent.
- A trial consists of a sequence of episodes, where at some point novelty is introduced and persists for the remainder of the episodes in a trial. For SmartEnv, each episode is a day in the life of the smart environment, and each episode consists of numerous sensor events that the agent is expected to classify as one of several activities.
- CDT (correctly determined trial) refers to a trial in which the agent detects novelty at some point after, but not before, novelty is introduced.
- $P_{Post,\alpha}$ = performance of target agent post-novelty.
- $P_{Post,\beta}$ = performance of baseline agent post-novelty.
- $P_{Pre,\beta}$ = performance of baseline agent pre-novelty.
- N_{pre} = number of episodes before novelty.
- N_{post} = number of episodes after novelty.
- m = number of episodes consider initial reaction after novelty or final reaction well after novelty at the end of a trial. Typically 5% of the total number of episodes in a trial.